Análise de empresas com base nos Dados Abertos do Governo Federal

Parte 2

Luiz Flavio Pereira

Linkedin | GitHub | Medium | Kaggle

Introdução

Esta segunda parte do trabalho tem por finalidade gerar um arquivo *.csv que mescla os dados obtidos na parte 1 deste projeto (Disponível aqui), com as informações específicas de cada empresa. Após a seleção, filtragem, tratamento dos dados e junção das tabelas, será gerado um arquivo csv combinado com os dados dos estabelecimentos e informações da empresa, que será utilizado na parte 3 deste projeto.

Execução da segunda parte do projeto

Para os dados contidos nesta categoria não é possível a filtragem por município como realizado na parte 1 deste projeto, recurso que reduz drasticamente o tamanho do dataframe gerado. Assim, todo o dataset deve lido e convertido em dataframe.

Devido ao enorme tamanho dos arquivos de dados das empresas e devido ao fato de ser necessário fazer o cruzamento e junção da tabela obtida na 1ª parte do trabalho com os arquivos descompactados, que totaliza mais de 50,4 milhões de linhas, não foi possível a execução desta tarefa com a biblioteca Pandas, uma vez que a memória disponibilizada pelo Google Colab não foi suficiente para a rotina. Assim sendo foi-se optado pela utilização da engine Spark através da API PySpark.

As principais etapas da 2ª parte são descritas abaixo:

- obtenção dos dados relativos às EMPRESAS;
- remoção das features desnecessárias ao escopo deste projeto;

- junção com os dados obtidos na parte 1 do projeto;
- tratamento dos dados;
- geração de arquivo de dados com as informações das empresas do município de interesse.

O notebook do projeto pode ser acessado através do GitHub.

Importante

Por conveniência o projeto foi todo realizado no ambiente de desenvolvimento `Google Colab` (memória disponibilizada e possibilidade de se utilizar os arquivos diretamente do Google Drive). Entretanto, o mesmo pode ser executado em qualquer outro ambiente de desenvolvimento desde que as bibliotecas, recursos e a base de dados necessários à execução estejam disponíveis.

In [ ]:

1. Setup do Spark

As etapas abaixo se fazem necessárias para a execução do PySpark no ambiente do Google Colab.

In [1]:
```
# Atualização do ambiente do Google Colab

!apt-get update
```

```
Get:1 https://cloud.r-project.org/bin/linux/ubuntu bionic-cran40/ InRelease [3,626 B]
Hit:2 http://archive.ubuntu.com/ubuntu bionic InRelease
Get:3 http://archive.ubuntu.com/ubuntu bionic-updates InRelease [88.7 kB]
Get:4 http://archive.ubuntu.com/ubuntu bionic-backports InRelease [74.6 kB]
Get:5 http://security.ubuntu.com/ubuntu bionic-security InRelease [88.7 kB]
Hit:6 http://ppa.launchpad.net/c2d4u.team/c2d4u4.0+/ubuntu bionic InRelease
Ign:7 https://developer.download.nvidia.com/compute/machine-learning/repos/ubuntu1804/x86_64  InRelease
Hit:8 https://developer.download.nvidia.com/compute/cuda/repos/ubuntu1804/x86_64  InRelease
Hit:9 https://developer.download.nvidia.com/compute/machine-learning/repos/ubuntu1804/x86_64  Release
Hit:10 http://ppa.launchpad.net/cran/libgit2/ubuntu bionic InRelease
Hit:11 http://ppa.launchpad.net/deadsnakes/ppa/ubuntu bionic InRelease
Get:12 http://archive.ubuntu.com/ubuntu bionic-updates/main amd64 Packages [3,397 kB]
Get:13 http://archive.ubuntu.com/ubuntu bionic-updates/universe amd64 Packages [2,318 kB]
Hit:14 http://ppa.launchpad.net/graphics-drivers/ppa/ubuntu bionic InRelease
Get:16 http://security.ubuntu.com/ubuntu bionic-security/universe amd64 Packages [1,540 kB]
Get:17 http://security.ubuntu.com/ubuntu bionic-security/main amd64 Packages [2,965 kB]
Fetched 10.5 MB in 2s (4,999 kB/s)
Reading package lists... Done
```

In [2]:
```
# Instalação do Spark no ambiente do Google Colab

!apt-get install openjdk-8-jdk-headless -qq > /dev/null
!wget -q https://archive.apache.org/dist/spark/spark-3.2.0/spark-3.2.0-bin-hadoop3.2.tgz
!tar xf spark-3.2.0-bin-hadoop3.2.tgz
!pip install -q findspark
```

In [3]:
```
# Importacao das bibliotecas do Spark e da API PySpark

import findspark
findspark.init('spark-3.2.0-bin-hadoop3.2')

import pyspark
sc = pyspark.SparkContext(appName='cnpj_pyspark')

from pyspark.sql import SparkSession
spark = SparkSession.builder.getOrCreate()

from pyspark.sql.functions import round, col
from pyspark.sql.types import StructType, StringType
```

---

2. Setup do projeto

In [4]:
```
# Importacao das bibliotecas necessarias a execucao das rotinas

import pandas as pd
from zipfile import ZipFile
```

In [5]:
```
# Remoção dos avisos de "warning" durante a execucao do programa

import warnings
warnings.filterwarnings("ignore")
```

In [6]:
```
# Montagem do Google Drive no ambiente do Google Colab

from google.colab import drive
drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

In [7]:
```
# Endereco do pasta com os arquivos do projeto no Google Drive
# Importante: o caminho das pastas devera ser alterado conforme o drive do usuario

drive_path = '/content/drive/MyDrive/cnpj_dados/'
```

In [8]:
```
# Lista para iteracao e descompactacao dos arquivos no ambiente do Google Colab

zip_files = ['Empresas0.zip', 'Empresas1.zip', 'Empresas2.zip', 'Empresas3.zip', 'Empresas4.zip',
             'Empresas5.zip', 'Empresas6.zip', 'Empresas7.zip', 'Empresas8.zip', 'Empresas9.zip']
```

In [9]:
```
# Descompactacao dos arquivos

for zip_file in zip_files:
    path = drive_path + zip_file

    with ZipFile(path, mode="r") as archive:
        archive.extractall()

    print(f'O arquivo {zip_file} foi descompactado.')
```

```
O arquivo Empresas0.zip foi descompactado.
O arquivo Empresas1.zip foi descompactado.
O arquivo Empresas2.zip foi descompactado.
O arquivo Empresas3.zip foi descompactado.
O arquivo Empresas4.zip foi descompactado.
O arquivo Empresas5.zip foi descompactado.
O arquivo Empresas6.zip foi descompactado.
O arquivo Empresas7.zip foi descompactado.
O arquivo Empresas8.zip foi descompactado.
O arquivo Empresas9.zip foi descompactado.
```

In [10]:
```
# Arquivos constantes no ambiente do Google Colab

! ls
```

```
drive                           K3241.K03200Y6.D20709.EMPRECSV
K3241.K03200Y0.D20709.EMPRECSV  K3241.K03200Y7.D20709.EMPRECSV
K3241.K03200Y1.D20709.EMPRECSV  K3241.K03200Y8.D20709.EMPRECSV
K3241.K03200Y2.D20709.EMPRECSV  K3241.K03200Y9.D20709.EMPRECSV
K3241.K03200Y3.D20709.EMPRECSV  sample_data
K3241.K03200Y4.D20709.EMPRECSV  spark-3.2.0-bin-hadoop3.2
K3241.K03200Y5.D20709.EMPRECSV  spark-3.2.0-bin-hadoop3.2.tgz
```

In [11]:
```
# Lista dos nomes dos arquivos descompactados para iteracao e geracao do dataframe com os dados das empresas

csv_files = ['K3241.K03200Y0.D20709.EMPRECSV', 'K3241.K03200Y1.D20709.EMPRECSV', 'K3241.K03200Y2.D20709.EMPRECS
             'K3241.K03200Y4.D20709.EMPRECSV', 'K3241.K03200Y5.D20709.EMPRECSV', 'K3241.K03200Y6.D20709.EMPRECS
             'K3241.K03200Y8.D20709.EMPRECSV', 'K3241.K03200Y9.D20709.EMPRECSV']
```

3. Geração do dataframe do dataset das Empresas

A documentação contendo os *layouts* dos arquivos disponibilizados pode ser encontrada neste link. De qualquer forma, as tabelas abaixo apresentam uma breve descrição das características contidas nos dados referentes apenas aos estabelecimentos e a justificativa para utilização ou não das *features* neste trabalho.

Campo	Descrição
CNPJ BÁSICO	NÚMERO BASE DE INSCRIÇÃO NO CNPJ (OITO PRIMEIROS DÍGITOS DO CNPJ).
RAZÃO SOCIAL/NOME EMPRESARIAL	NOME EMPRESARIAL DA PESSOA JURÍDICA

Campo	Descrição
NATUREZA JURÍDICA	CÓDIGO DA NATUREZA JURÍDICA
QUALIFICAÇÃO RESPONSÁVEL	QUALIFICAÇÃO DA PESSOA FÍSICA RESPONSÁVEL PELA EMPRESA
CAPITAL SOCIAL DA EMPRESA	CAPITAL SOCIAL DA EMPRESA
PORTE DA EMPRESA	CÓDIGO DO PORTE DA EMPRESA: 00. NÃO INFORMADO; 01. MICRO EMPRESA; 03. EMPRESA DE PEQUENO PORTE; 05. DEMAIS
ENTE FEDERATIVO RESPONSÁVEL	O ENTE FEDERATIVO RESPONSÁVEL É PREENCHIDO PARA OS CASOS DE ÓRGÃOS E ENTIDADES DO GRUPO DE NATUREZA JURÍDICA 1XXX. PARA AS DEMAIS NATUREZAS, ESTE ATRIBUTO FICA EM BRANCO.

In [11]:

Após análise preliminar dos dados e pautado pela motivação do projeto foram selecionadas as seguintes colunas conforme descrito abaixo. Para não haver perda de informações, mesmo as variáveis numéricas serão importadas como strings.

Indíce	Campo	Tipo de variável	Utilizada?	Justificativa
0	CNPJ BÁSICO	Objeto	Sim	Identificação da empresa
1	RAZÃO SOCIAL/NOME EMPRESARIAL	Objeto	Sim	Identificação da empresa
2	NATUREZA JURÍDICA	Objeto	Sim	Define a natureza jurídica da empresa
3	QUALIFICAÇÃO RESPONSÁVEL	Objeto	Não	Item não considerado relevante para esta análise
4	CAPITAL SOCIAL DA EMPRESA	Objeto	Sim	Define o capital social da empresa e pode indicar o porte da mesma
5	PORTE DA EMPRESA	Objeto	Sim	Define o porte da empresa e tipo de cadastro
6	ENTE FEDERATIVO RESPONSÁVEL	Objeto	Não	Item não considerado relevante para esta análise

In [11]:

In [12]:
```
# Definicao do schema do dataframe do PySpark para importacao dos dados dos datasets

schema = StructType().add("cnpj_basico_2", StringType(),True) \
                     .add("razao_social", StringType(),True) \
                     .add("natureza_juridica", StringType(),True) \
                     .add("qualificacao", StringType(),True) \
                     .add("capital_social", StringType(),True) \
                     .add("porte_empresa", StringType(),True) \
                     .add("ente_federativo", StringType(),True)
```

In [13]:
```
# Esta rotina efetua a leitura dos 10 arquivos csv das empresas, descompactados previamente
# um por vez, e os armazena em um dataframe, que sera utilizado no restante do codigo

for csv in csv_files:
    if csv == 'K3241.K03200Y0.D20709.EMPRECSV':
        df = spark.read.csv(csv, nullValue='NA', sep=';', schema=schema)
        print(f'O dataset {csv} possui {df.count()} linhas.')
    else:
        df2 = spark.read.csv(csv, nullValue='NA', sep=';', schema=schema)
        print(f'O dataset {csv} possui {df2.count()} linhas.')

        df = df.unionByName(df2)
```

```
O dataset K3241.K03200Y0.D20709.EMPRECSV possui 10011164 linhas.
O dataset K3241.K03200Y1.D20709.EMPRECSV possui 4494860 linhas.
O dataset K3241.K03200Y2.D20709.EMPRECSV possui 4494860 linhas.
O dataset K3241.K03200Y3.D20709.EMPRECSV possui 4494860 linhas.
O dataset K3241.K03200Y4.D20709.EMPRECSV possui 4494860 linhas.
O dataset K3241.K03200Y5.D20709.EMPRECSV possui 4494860 linhas.
O dataset K3241.K03200Y6.D20709.EMPRECSV possui 4494860 linhas.
O dataset K3241.K03200Y7.D20709.EMPRECSV possui 4494860 linhas.
O dataset K3241.K03200Y8.D20709.EMPRECSV possui 4494860 linhas.
O dataset K3241.K03200Y9.D20709.EMPRECSV possui 4494860 linhas.
```

In [14]:
```
# Apresenta a forma do dataframe

print(f'O dataframe df possui {df.count()} linhas e {len(df.columns)} colunas')
```

```
O dataframe df possui 50464904 linhas e 7 colunas
```

In [15]:
```
# Exibe as 10 primeiras linhas do dataframe
```

```
df.show(10)
```

```
+-------------+--------------------+-----------------+------------+--------------+-------------+--------------
-+
|cnpj_basico_2|        razao_social|natureza_juridica|qualificacao|capital_social|porte_empresa|ente_federativ
o|
+-------------+--------------------+-----------------+------------+--------------+-------------+--------------
-+
|     41273600|AVANILSON BRUNO M...|             2135|          50|      50000,00|           01|
|
|     41273601|GABRIELA HELENA F...|             2135|          50|       2000,00|           01|
|
|     41273602|FABIO SOUZA DO RO...|             2135|          50|      15000,00|           01|
|
|     41273603|GRAFLINE ACESSORI...|             2062|          49|      10000,00|           01|
|
|     41273604|RUMO - ESTUDIO DE...|             2062|          49|      10000,00|           01|
|
|     41273605|WALLACE DE OLIVEI...|             2135|          50|       1000,00|           01|
|
|     41273606|MARCOS CESAR DE M...|             2135|          50|      72000,00|           01|
|
|     41273607|LAYANE SCARLETT D...|             2135|          50|          1,00|           01|
|
|     41273608|FRANCISCA SAMPAIO...|             2135|          50|          0,00|           01|
|
|     41273609|INGRID DIAS ALVES...|             2135|          50|      10000,00|           01|
|
+-------------+--------------------+-----------------+------------+--------------+-------------+--------------
-+
only showing top 10 rows
```

In [16]:
```
# Exibe o schema do Dataframe

df.printSchema()
```

```
root
 |-- cnpj_basico_2: string (nullable = true)
 |-- razao_social: string (nullable = true)
 |-- natureza_juridica: string (nullable = true)
 |-- qualificacao: string (nullable = true)
 |-- capital_social: string (nullable = true)
 |-- porte_empresa: string (nullable = true)
 |-- ente_federativo: string (nullable = true)
```

In [17]:
```
# Remocao das colunas qualificacao e ente_federativo

df_emp = df.select(['cnpj_basico_2', 'razao_social', 'natureza_juridica', 'capital_social', 'porte_empresa'])
df_emp.show(10)
```

```
+-------------+--------------------+-----------------+--------------+-------------+
|cnpj_basico_2|        razao_social|natureza_juridica|capital_social|porte_empresa|
+-------------+--------------------+-----------------+--------------+-------------+
|     41273600|AVANILSON BRUNO M...|             2135|      50000,00|           01|
|     41273601|GABRIELA HELENA F...|             2135|       2000,00|           01|
|     41273602|FABIO SOUZA DO RO...|             2135|      15000,00|           01|
|     41273603|GRAFLINE ACESSORI...|             2062|      10000,00|           01|
|     41273604|RUMO - ESTUDIO DE...|             2062|      10000,00|           01|
|     41273605|WALLACE DE OLIVEI...|             2135|       1000,00|           01|
|     41273606|MARCOS CESAR DE M...|             2135|      72000,00|           01|
|     41273607|LAYANE SCARLETT D...|             2135|          1,00|           01|
|     41273608|FRANCISCA SAMPAIO...|             2135|          0,00|           01|
|     41273609|INGRID DIAS ALVES...|             2135|      10000,00|           01|
+-------------+--------------------+-----------------+--------------+-------------+
only showing top 10 rows
```

4. Geração do dataframe dos dados obtidos da parte 1 do projeto (dados de Estabelecimento)

In [18]:
```
# Define o endereco do arquivo csv gerado na parte 1 deste projeto

csv_parte_1 = drive_path + 'df_municipio_pocos_de_caldas.csv'
```

In [19]:
```
# Leitura do arquivo gerado na parte 1 do projeto e geracao de dataframe com os dados deste
```

```
# Exibicao do schema do dataframe

df_parte1 = spark.read.csv(csv_parte_1, nullValue='NA', sep=',', header=True)
df_parte1.printSchema()
```

```
root
 |-- cnpj_basico: string (nullable = true)
 |-- cnpj_ordem: string (nullable = true)
 |-- cnpj_dv: string (nullable = true)
 |-- identificador_matriz_filial: string (nullable = true)
 |-- nome_fantasia: string (nullable = true)
 |-- situacao_cadastral: string (nullable = true)
 |-- data_situacao_cadastral: string (nullable = true)
 |-- motivo_situacao_cadastral: string (nullable = true)
 |-- data_de_inicio_atividade: string (nullable = true)
 |-- cnae_fiscal_principal: string (nullable = true)
 |-- numero: string (nullable = true)
 |-- complemento: string (nullable = true)
 |-- cep: string (nullable = true)
```

In [22]:
```
# Exibe as 10 primeiras linhas do dataframe

df_parte1.show(10)
```

```
+-----------+----------+-------+---------------------------+--------------------+------------------+----------
-------------+-------------------------+-----------------------+--------------------+------+-----------+----
----+
|cnpj_basico|cnpj_ordem|cnpj_dv|identificador_matriz_filial|       nome_fantasia|situacao_cadastral|data_situa
cao_cadastral|motivo_situacao_cadastral|data_de_inicio_atividade|cnae_fiscal_principal|numero|complemento|
cep|
+-----------+----------+-------+---------------------------+--------------------+------------------+----------
-------------+-------------------------+-----------------------+--------------------+------+-----------+----
----+
|   02951824|      0002|     53|                          2|                null|                04|
20190124|                       63|               20021111|             4729699|  1248|          A|37701386|
|   02114490|      0002|     63|                          2|SOLANGE MODA E AC...|                08|
20101109|                       01|               20011016|             4781400|    15|       null|37701010|
|   03907506|      0001|     77|                          1|                null|                02|
20051103|                       00|               20000626|             8650003|   221|   CONJ  23|37701025|
|   05571365|      0001|     90|                          1|                null|                08|
20130415|                       01|               20030331|             4759899|   404|       null|37701033|
|   03913359|      0001|     48|                          1|           SONHO MEU|                08|
20110426|                       01|               20000623|             4729699|   255|       null|37706147|
|   00604792|      0001|     03|                          1|                null|                08|
20131209|                       01|               19950510|             5611203|   364|       null|37701386|
|   05577240|      0001|     77|                          1|                null|                04|
20180925|                       63|               20030319|             4781400|   606|       null|37701094|
|   02226925|      0001|     80|                          1|  REQUINTE FOLHEADOS|                08|
20081231|                       71|               19971103|             4649408|   271|       null|37701010|
|   00607194|      0001|     98|                          1|                null|                02|
20051103|                       00|               19950518|             8650099|   125|       null|37701025|
|   00609668|      0001|     30|                          1|ALICE PRESENTES E...|                08|
20051219|                       01|               19950517|             4713004|   554|    LOJA 06|37701016|
+-----------+----------+-------+---------------------------+--------------------+------------------+----------
-------------+-------------------------+-----------------------+--------------------+------+-----------+----
----+
only showing top 10 rows
```

In [23]:
```
# Apresenta a forma do dataframe da parte 1

print(f'O dataframe df_municipio contem {df_parte1.count()} linhas e {len(df_parte1.columns)} colunas.')
```

```
O dataframe df_municipio contem 54603 linhas e 13 colunas.
```

5. Junção dos dataframes

In [24]:
```
# Juncao do dataframe com os dados das empresas com o dataframe da parte 1 do projeto
# atraves do dado de CNPJ, identificador unico das empresas

# joined = df_pocos.join(df, df_pocos.cnpj_basico == df._c0, how="inner")
# joined.show()

df_emp.createOrReplaceTempView("DF")
df_parte1.createOrReplaceTempView("DF_POCOS")

df_emp_parte1 = spark.sql("SELECT * FROM DF_POCOS AS P \
                           LEFT JOIN DF AS D \
```

```
                        ON P.cnpj_basico = D.cnpj_basico_2")

df_emp_parte1.show()
```

```
+----------+----------+-------+--------------------------+-------------------+-----------------+----------
------------+------------------------+----------------------+--------------------+------+---------------
--+--------+------------+-------------------+----------------+-------------+-------------+
|cnpj_basico|cnpj_ordem|cnpj_dv|identificador_matriz_filial|      nome_fantasia|situacao_cadastral|data_situa
cao_cadastral|motivo_situacao_cadastral|data_de_inicio_atividade|cnae_fiscal_principal|numero|      complemen
to|     cep|cnpj_basico_2|        razao_social|natureza_juridica|capital_social|porte_empresa|
+----------+----------+-------+--------------------------+-------------------+-----------------+----------
------------+------------------------+----------------------+--------------------+------+---------------
--+--------+------------+-------------------+----------------+-------------+-------------+
|   37318064|      0001|     45|                         1|            AMARELO|               02|
20200604|                        00|              20200604|             4772500|   325|             CONJ|37
701561|     37318064|FERNANDA PAULA JA...|              2135|       1000,00|           01|
|   37362540|      0001|     25|                         1|           REVIGORE|               02|
20200609|                        00|              20200609|             9602502|   314|          CASA  2|37
701038|     37362540|REVIGORE CLINICA ...|              2062|      40000,00|           01|
|   37313704|      0001|     24|                         1|               null|               02|
20200603|                        00|              20200603|             8112500|  1061|             null|37
701714|     37313704|ASSOCIACAO PROCON...|              3999|          0,00|           05|
|   02232252|      0001|     71|                         1|               null|               04|
20180911|                        63|              19971111|             4617600|    73|          SALA 22|37
701275|     02232252|REPRESENTACAO DAN...|              2062|          0,00|           05|
|   00609668|      0001|     30|                         1|ALICE PRESENTES E...|               08|
20051219|                        01|              19950517|             4713004|   554|          LOJA 06|37
701016|     00609668|COMERCIAL MURARO ...|              2062|      10000,00|           01|
|   37336785|      0001|     88|                         1|      ZARREF MARTINS|               02|
20200605|                        00|              20200605|             8230001|    06|             null|37
710416|     37336785|LUCAS FERRAZ MART...|              2135|       6000,00|           01|
|   37392389|      0001|     78|                         1|               null|               08|
20220621|                        01|              20200612|             4754702|   180|             null|37
706283|     37392389|SILVIO ROBERTO DA...|              2135|       1000,00|           01|
|   05577240|      0001|     77|                         1|               null|               04|
20180925|                        63|              20030319|             4781400|   606|             null|37
701094|     05577240|NILCEA APARECIDA ...|              2135|          0,00|           01|
|   37351222|      0001|     69|                         1|TRANSPORTE POR AP...|               08|
20211004|                        01|              20200608|             5229099|    06|             null|37
704166|     37351222|JOSE GERSINO PERE...|              2135|       3000,00|           01|
|   26945321|      0001|     52|                         1|             AGROVET|               02|
20170125|                        00|              20170125|             4789004|   380|LOJA  01 / 02 / 03|37
706147|     26945321|             AGROVET|              2135|      20000,00|           01|
|   03937674|      0001|     05|                         1|      JD INFORMATICA|               08|
20081231|                        71|              20000714|             4751201|   511|             null|37
701004|     03937674|J & D INFORMATICA...|              2062|          0,00|           05|
|   37393412|      0001|     49|                         1|SANDERCEIA VITORI...|               02|
20200612|                        00|              20200612|             4321500|   356|             CONJ|37
701445|     37393412|SANDERCEIA VITORI...|              2135|       1000,00|           01|
|   37445593|      0001|     00|                         1|        PR CONFECCAO|               02|
20200617|                        00|              20200617|             1413403|    98|             BRCAO|37
704016|     37445593|ALIFER PIRES PELE...|              2135|          1,00|           01|
|   03926356|      0001|     49|                         1|          AGUA VIVA|               04|
20180918|                        63|              20000710|             1412601|   220|             null|37
701099|     03926356| BMB CONFECCOES LTDA|              2062|          0,00|           01|
|   37458107|      0001|     98|                         1|TWJ GESTAO ADMINI...|               02|
20200618|                        00|              20200618|             8211300|   774|          CASA  19|37
703263|     37458107|TWJ GESTAO ADMINI...|              2062|       7000,00|           01|
|   03907506|      0001|     77|                         1|               null|               02|
20051103|                        00|              20000626|             8650003|   221|          CONJ  23|37
701025|     03907506|CLINICA DE PSICOL...|              2240|      15000,00|           01|
|   04531266|      0003|     83|                         2|               null|               08|
20061219|                        01|              20030707|             4784900|   198|             null|37
704284|     04531266|          BIGAS LTDA|              2062|       5000,00|           01|
|   03913359|      0001|     48|                         1|           SONHO MEU|               08|
20110426|                        01|              20000623|             4729699|   255|             null|37
706147|     03913359|CARVALHO & DIONIS...|              2062|          0,00|           01|
|   00604792|      0001|     03|                         1|               null|               08|
20131209|                        01|              19950510|             5611203|   364|             null|37
701386|     00604792|LANCHAR LANCHES M...|              2062|       3000,00|           01|
|   37453819|      0001|     14|                         1|               null|               02|
20181003|                        00|              20181003|             8112500|   180|             null|37
705614|     37453819|CONDOMINIO RESIDE...|              3085|          0,00|           05|
+----------+----------+-------+--------------------------+-------------------+-----------------+----------
------------+------------------------+----------------------+--------------------+------+---------------
--+--------+------------+-------------------+----------------+-------------+-------------+
only showing top 20 rows
```

In [25]:
```
# Apresenta a forma do dataframe combinado
```

```
print(f'O dataframe apos o join contem {df_emp_parte1.count()} linhas e {len(df_emp_parte1.columns)} colunas.'
```

```
O dataframe apos o join contem 54603 linhas e 18 colunas.
```

In [26]:
```
# Apresenta o schema do dataframe combinado

df_emp_parte1.printSchema()
```

```
root
 |-- cnpj_basico: string (nullable = true)
 |-- cnpj_ordem: string (nullable = true)
 |-- cnpj_dv: string (nullable = true)
 |-- identificador_matriz_filial: string (nullable = true)
 |-- nome_fantasia: string (nullable = true)
 |-- situacao_cadastral: string (nullable = true)
 |-- data_situacao_cadastral: string (nullable = true)
 |-- motivo_situacao_cadastral: string (nullable = true)
 |-- data_de_inicio_atividade: string (nullable = true)
 |-- cnae_fiscal_principal: string (nullable = true)
 |-- numero: string (nullable = true)
 |-- complemento: string (nullable = true)
 |-- cep: string (nullable = true)
 |-- cnpj_basico_2: string (nullable = true)
 |-- razao_social: string (nullable = true)
 |-- natureza_juridica: string (nullable = true)
 |-- capital_social: string (nullable = true)
 |-- porte_empresa: string (nullable = true)
```

In [27]:
```
# Remocao da coluna cnpj_basico_2 (redundante)

# df_joined = df_joined.drop(*['cnpj_basico_2'])
df_emp_parte1 = df_emp_parte1.drop(df_emp_parte1.cnpj_basico_2)
df_emp_parte1.show(10)
```

```
+-----------+----------+-------+---------------------------+--------------------+------------------+----------
-------------+-------------------------+-----------------------+--------------------+------+-----------+----
----+--------------------+----------------+-------------+-------------+
|cnpj_basico|cnpj_ordem|cnpj_dv|identificador_matriz_filial|       nome_fantasia|situacao_cadastral|data_situa
cao_cadastral|motivo_situacao_cadastral|data_de_inicio_atividade|cnae_fiscal_principal|numero|complemento|
cep|        razao_social|natureza_juridica|capital_social|porte_empresa|
+-----------+----------+-------+---------------------------+--------------------+------------------+----------
-------------+-------------------------+-----------------------+--------------------+------+-----------+----
----+--------------------+----------------+-------------+-------------+
|   37318064|      0001|     45|                          1|             AMARELO|                02|
20200604|                       00|               20200604|             4772500|   325|       CONJ|37701561|
FERNANDA PAULA JA...|             2135|       1000,00|           01|
|   37362540|      0001|     25|                          1|            REVIGORE|                02|
20200609|                       00|               20200609|             9602502|   314|    CASA  2|37701038|
REVIGORE CLINICA ...|             2062|      40000,00|           01|
|   37313704|      0001|     24|                          1|                null|                02|
20200603|                       00|               20200603|             8112500|  1061|       null|37701714|
ASSOCIACAO PROCON...|             3999|          0,00|           05|
|   00609668|      0001|     30|                          1|ALICE PRESENTES E...|                08|
20051219|                       01|               19950517|             4713004|   554|    LOJA 06|37701016|
COMERCIAL MURARO ...|             2062|      10000,00|           01|
|   37336785|      0001|     88|                          1|      ZARREF MARTINS|                02|
20200605|                       00|               20200605|             8230001|    06|       null|37710416|
LUCAS FERRAZ MART...|             2135|       6000,00|           01|
|   37392389|      0001|     78|                          1|                null|                08|
20220621|                       01|               20200612|             4754702|   180|       null|37706283|
SILVIO ROBERTO DA...|             2135|       1000,00|           01|
|   05577240|      0001|     77|                          1|                null|                04|
20180925|                       63|               20030319|             4781400|   606|       null|37701094|
NILCEA APARECIDA ...|             2135|          0,00|           01|
|   37351222|      0001|     69|                          1|TRANSPORTE POR AP...|                08|
20211004|                       01|               20200608|             5229099|    06|       null|37704166|
JOSE GERSINO PERE...|             2135|       3000,00|           01|
|   03907506|      0001|     77|                          1|                null|                02|
20051103|                       00|               20000626|             8650003|   221|   CONJ  23|37701025|
CLINICA DE PSICOL...|             2240|      15000,00|           01|
|   03913359|      0001|     48|                          1|           SONHO MEU|                08|
20110426|                       01|               20000623|             4729699|   255|       null|37706147|
CARVALHO & DIONIS...|             2062|          0,00|           01|
+-----------+----------+-------+---------------------------+--------------------+------------------+----------
-------------+-------------------------+-----------------------+--------------------+------+-----------+----
----+--------------------+----------------+-------------+-------------+
only showing top 10 rows
```

In [28]:
```
# Converte o dataframe PySpark para um dataframe formato Pandas

df_emp_parte1_pd = df_emp_parte1.toPandas()
```

In [29]:
```
# Apresenta informacoes do dataframe Pandas

df_emp_parte1_pd.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 54603 entries, 0 to 54602
Data columns (total 17 columns):
 #   Column                        Non-Null Count  Dtype 
---  ------                        --------------  ----- 
 0   cnpj_basico                   54603 non-null  object
 1   cnpj_ordem                    54603 non-null  object
 2   cnpj_dv                       54603 non-null  object
 3   identificador_matriz_filial   54603 non-null  object
 4   nome_fantasia                 34431 non-null  object
 5   situacao_cadastral            54603 non-null  object
 6   data_situacao_cadastral       54603 non-null  object
 7   motivo_situacao_cadastral     54603 non-null  object
 8   data_de_inicio_atividade      54603 non-null  object
 9   cnae_fiscal_principal         54603 non-null  object
 10  numero                        54603 non-null  object
 11  complemento                   18056 non-null  object
 12  cep                           54603 non-null  object
 13  razao_social                  54603 non-null  object
 14  natureza_juridica             54603 non-null  object
 15  capital_social                54603 non-null  object
 16  porte_empresa                 54603 non-null  object
dtypes: object(17)
memory usage: 7.1+ MB
```

In [30]:
```
# Exibe as 5 primieras linhas do dataframe Pandas

df_emp_parte1_pd.head()
```

Out[30]:

	cnpj_basico	cnpj_ordem	cnpj_dv	identificador_matriz_filial	nome_fantasia	situacao_cadastral	data_situacao_cadastral	motivo_situacao
0	00065369	0001	82	1	None	02	20050827	
1	00328348	0001	02	1	None	04	20190124	
2	00331566	0001	04	1	None	08	20021107	
3	00354051	0001	11	1	KLIKOS CALCADOS	08	20190612	
4	00365630	0001	60	1	None	08	20081128	

In [31]:
```
# Substituicao dos caracteres " ; : das instancias para evitar conflitos na geracao do arquivo csv

df_emp_parte1_pd = df_emp_parte1_pd.replace('[";:;]',' ', regex=True)
```

6. Geração e exportação do arquivo CSV finalizado da parte 2

In [32]:
```
# Mascara para o nome do csv

mascara_nome_arquivo = 'df_pocos_pt2.csv'
```

In [33]:
```
# Criacao do arquivo csv

df_emp_parte1_pd.to_csv(mascara_nome_arquivo, index=False, encoding='utf-8', na_rep='', sep=';')
print(f'O arquivo {mascara_nome_arquivo} foi gerado com sucesso.')
```

```
O arquivo df_pocos_pt2.csv foi gerado com sucesso.
```

In [34]:
```
# Criacao uma copia do arquivo csv finalizado no diretorio de arquivos de CNPJ do Google Drive

! cp $mascara_nome_arquivo $drive_path
print(f'O arquivo {mascara_nome_arquivo} foi salvo na pasta {drive_path[15:]} do Google Drive.')
```

```
O arquivo df_pocos_pt2.csv foi salvo na pasta MyDrive/cnpj_dados/ do Google Drive.
```